UNIT-1

DATA MINING

Definition of Data Mining :-

"Data Mining" is the process of discovering / extracting interesting patterns or rules or constraints from the large amount of data. It is simply known as "Knowledge Discovery in Database" (KDD)

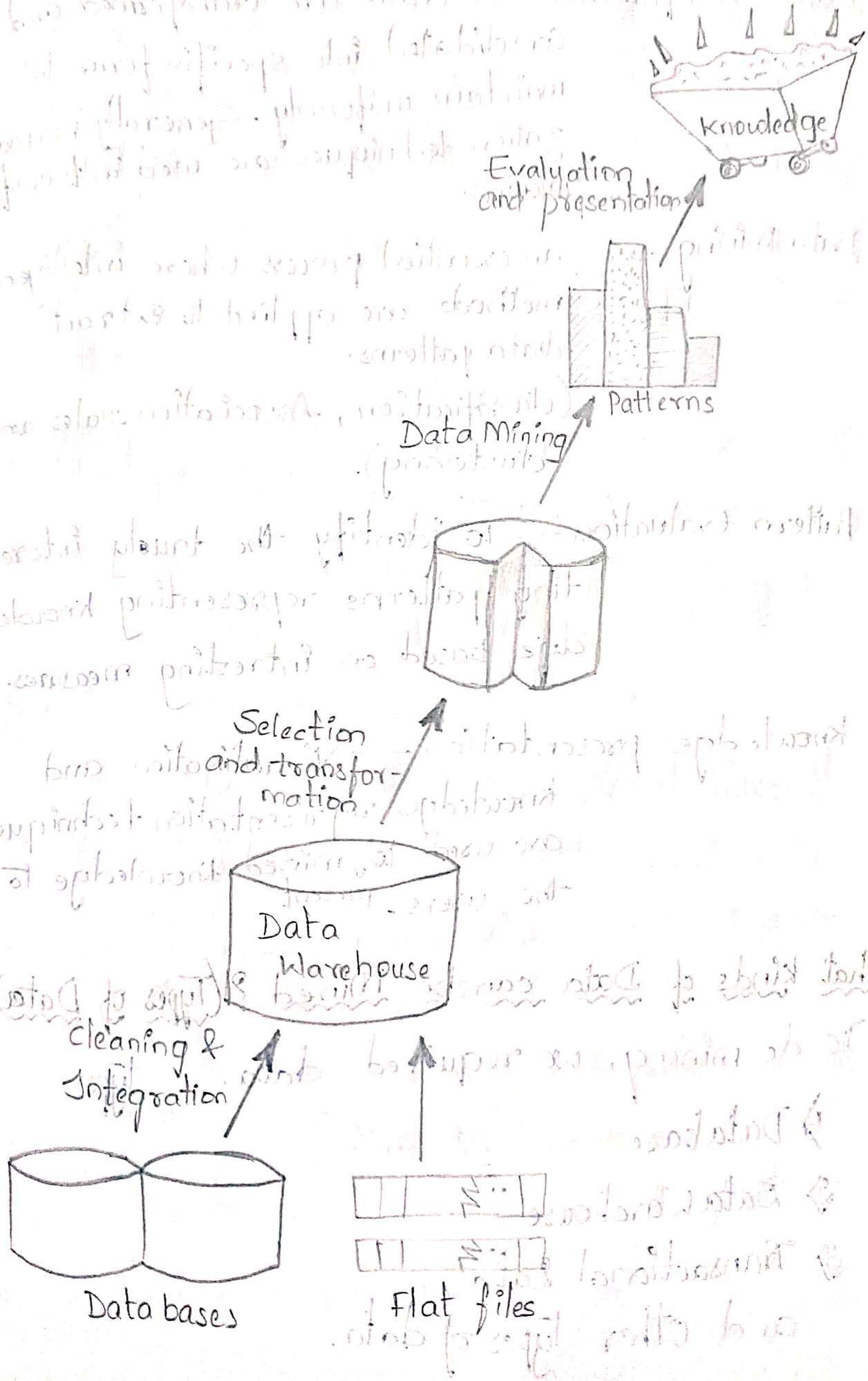
As a "knowledge Discovery process", it typically involves
- data cleaning
- data Integration
- data selection
- data transformation
- pattern discovery
- pattern evaluation
- knowledge presentation

KDD Process : The KDD process is an iterative sequence of the following steps.

(1) Data cleaning — to remove noise and inconsistent data. When the same data exists in different formats in multiple tables. (duplicate)

(2) Data Integration — Where multiple data sources / data bases may be combined. All these data bases are heterogenous type.

Figure:- Data Mining as a step in the process of knowledge discovery



③ Data selection —— relevant data is selected from the database for analysis task.

④ Data Transformation —— data are transformed and consolidated into specific form to maintain uniformly. Generally normalization techniques are used in transformation.

⑤ Data Mining —— an essential process where intelligent methods are applied to extract data patterns.
(classification, Association rules and clustering).

⑥ Pattern Evaluation — To identify the truely interesting patterns representing knowledge based on intresting measures.

⑦ knowledge presentation — Visualization and knowledge representation techniques are used to present mined knowledge to the users.

What kinds of Data can be Mined ? (Types of Data)

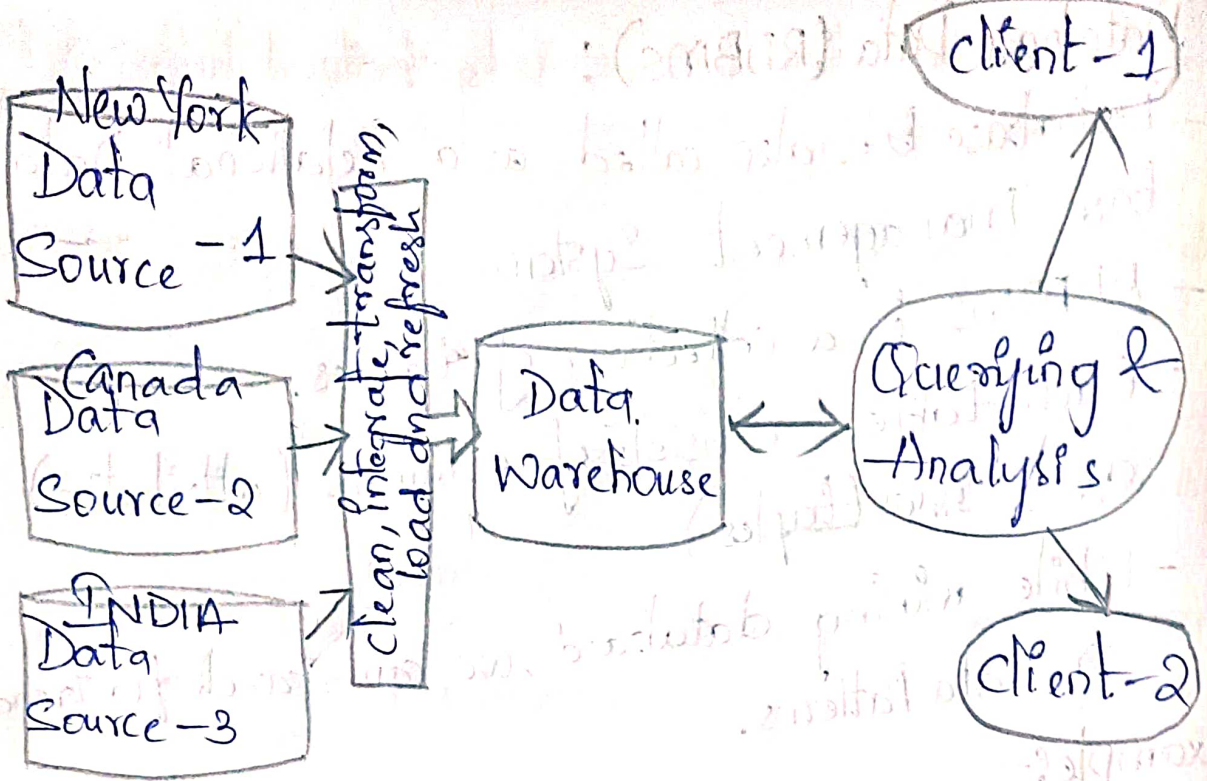
* To do mining, we required data. types

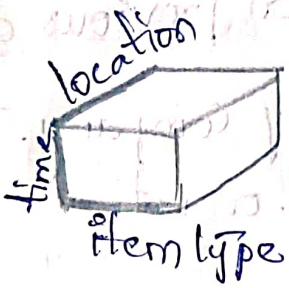
1) Database
2) Data Warehouse
3) Transactional Data
and other types of data.

(1) Database Data (RDBMS): Data is stored in form of Tables.
- Database is also called as a Relational Data Base Management System.
- RDBMS is a collection of tables.
- Each tables consists of columns (attributes) and rows (tuples)
- While mining database, we can search for trends or data Patterns. (set is the o/p)

Example:-
1. Analysing existing customer data to predict the credit risks of new customers. based on previous data.
2. Analysing sales data for any company. To check is there any deviations or not.

(2) Data Warehouse data:
- Data Warehouse is a collection of integrated data from different sources & with querying and decision making on data.. modifications
↑ sales; ↓ sales, profits
- In Data Warehouse, data is stored in multi-dimensional structure (data cube) where each dimension represents an attributes.





(3) Transactional Database:

- Each record in Transactional DB is a transac
  Such as Sales, Customers purchase, flight booking
  User clicks on web page. (click on Ads, banners, etc)
- Transaction include transaction_id, transaction name, transaction amt, etc. & list of other items making transactions.
- from transaction DB, we can mine frequent patterns.

Other Types of data?

- Sequence data (stock market related data), data streams (data which is continuously transmitted), Spatial data (maps),
- Engineering Design Data (Integreated circuits), hypertext, multimedia web data, etc.

Data Mining Functionalities:—

- There are five functionalities

1) class/concept Description: Characterization and Discrimination.
2) Mining frequent patterns, association and correlations.
3) Classification and Regression for Predictive Analysis.
4) Cluster Analysis
5) Outlier Analysis.

(1) Class/Concept Description: Characterization and discrimination:

- Data is always associated with class or concept.
- It can be useful to describe individual class or concept in summered, concise and precise in terms. Such description ~~is called as~~ of class

or Concept is called as a class/Concept description.

- This description can be done in 2-ways.

(i) Data characterisation:

* It refers to the Summarization of the class or concept of data.

Eg:- General overview

(ii) Data Discrimination:

* Is a comparision of the common features of the class or concept with other class/concepts

Eg:- Barcharts, Curves, etc.

(2) Mining frequent Patterns, associations and Correlations:

- frequent Patterns - The patterns that occurs frequently in data.
  - frequent itemsets (data items / data objects)
  - frequent subsequence
  - frequent substructure

- Association Analysis - It is a way of identifying the relation between various items.

- Eg: Association Analysis used to determine sales of items that are frequently purchased together

- Correlation Analysis - It is a mathematical technique which shows how strongly pair of attributes are related together

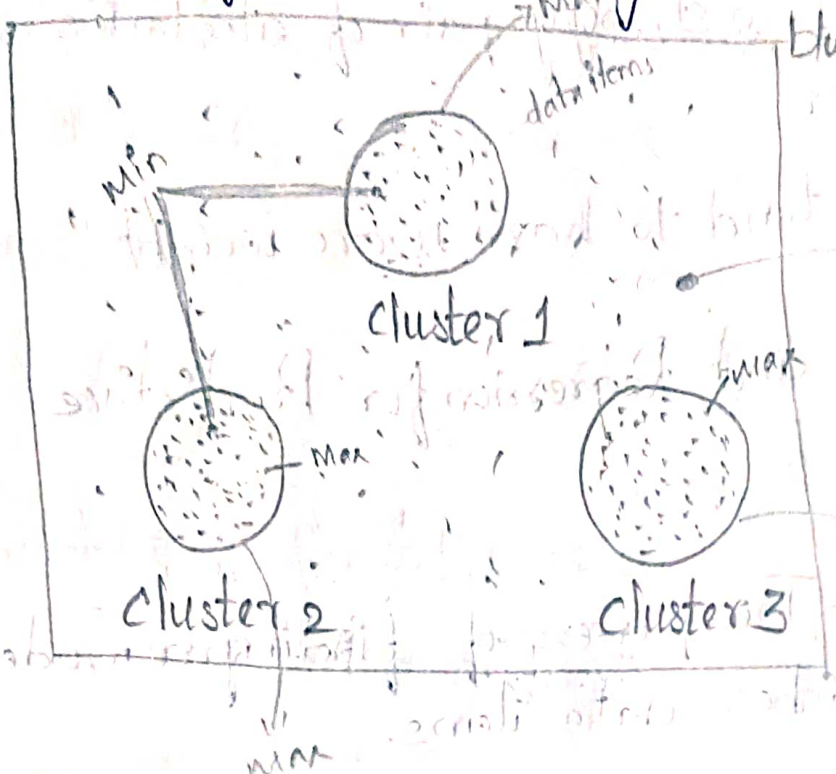
Eg:- Tall people (1) tend to have more (2) weight

(3) Classification and Regression for Predictive Analysis :- prediction of data — if any data is wrong predict data & fill missing data.

- Classification - The process of finding a model that distinguishes data items.
  - The decision tree is used for classification.
  - Decision Tree : A decision tree is a flow chart like structure, where each
    Node — denotes a test on an attribute value
    branch — represents outcome of the test
    tree leaves - Represents classes or class distribution
- Regression - Statistical methodology that is used for numeric prediction of missing data can be done based on previous data — Target data

(4) Cluster Analysis : → group
- The data items are clustered based on the principle of maximising the intraclass within same class

Similarity and minimising the interclass similarity btw two diff classes



- The Analysis of this above cluster is an cluster Analysis.

(5) Outlier Analysis:

- Is also known as anomaly mining. → which is not obeying the rule.
- Among the data items in a dataset, there may be some items which do not follow the general behaviour of data is known as outlier.
- Many data mining methods remove the outlier as noise & exceptions.
- The Analysis of outlier data is referred as Outlier Analysis & anomaly mining.

Eg:- {2, 4, 6, 7, 8, 10, 12, ...}
↓
odd no. — outlier

Intresting Patterns :-

- In a data mining system, every day millions of data patterns are generated
- Among all these patterns generated, how many are really interesting ? — How it is useful for users
  needed
- Actually, a small fraction of patterns generated would be of interest to any given user.

This raises (3) Questions

(1) What makes patterns interesting ?
A pattern is interesting if it is
- Easily understood by humans
- Valid on new/test data
- potentially useful for the user

(2) Can data mining system generate all of the interesting patterns ? — dress
- It refers to completeness of a data Mining system
- In reality it is not possible for a data Mining System to generate all interesting patterns.

(3) Can data Mining system generates only interesting patterns ?

It refers to optimization of a data mining System. (similar stress)

- In reality it is not possible for a data mining System to generate all interesting patterns.
- Generating only interesting patterns is a challenging.
- If only interesting patterns are generated, it becomes easy and efficient for the uses. (time is saved)

Classification of Data Mining Systems :-

Data Mining Systems are classified based on several criteria. They are:

(1) Classification based on mined database: based on type of database that is been mined.
   - Relational
   - Transactional
   - Object-relational
   - Data Warehouse

(2) Classification based on type of knowledge mined
   - Characterization
   - discrimination

- Association and Correlation analysis
- Classification
- Prediction
- Outlier Analysis
- Evolution Analysis

(3) Classification based on kinds of techniques used
- Machine Learning
- Statistics
- Neural Networks
- Pattern Recognition
- Data Warehouse Oriented techniques, etc.

(4) Classification based on applications adapted.
- Finance
- Telecommunications
- DNA
- Stock Markets
- Email etc

Data Mining Task Primitives :-

- A Data Mining task is represented in form of a data mining query is defined in terms of Data Mining task primitives.
- DM task primitives will allow the users to interactively communicate with the DM syster

- There are 5 data mining task Primitives.

(1) Set of task relevant data to be mined - It Specifies the portion of data where the user is interested

eg: Supermarket → fruits & Arrays
→ Age

It is also called as a relevant attributes.

(2) Specifies the kind of knowledge to be mined -
- Characterization & discrimination
- Association and Correlation Analysis
- Classification and Regression Analysis
- Cluster and Outlier Analysis.

(3) The background knowledge to be used in discovery process -
knowledge
based on domain we choose prog: lang - Java, Python...

- Concept hierarchies are used for the background knowledge in discovery process.

(4) The interestingness measures and thresholds for pattern Evaluation
(how much interest the user is showing on that pattern
threshold → unfit

- How useful the patterns are and how much interest the user is showing on that pattern.

- If the Interestingness measures value is less than specified threshold value that such patterns are considered as uninterest pattern.

(5). The Expected representation for visualizing the discovered patterns.
- The users are expected the representation of discovered patterns.
- The patterns can be rules/tables/patterns/charts/graphs.

Integration of Data Mining System with a Data Warehouse :—

- In this concept, we combine the Data Mining with Data Warehouse / Data Base. Because to establish the communication between the Data Mining & Data Warehouse.
- If there is no Integration, then there is no communication with Data Base / Data Warehouse

Four Integration Schema's :

(1) No Coupling : No integration with Data Mining and Data Warehouse.

- DM system will not use any functions of DW/DB. i.e there is no communication with DW/DB. (can't fetch data)
- In this scenario, it will communicate with other storage methods like file system.

(2) Loose Coupling: ~~20%~~ 100
- It will integrate only upto some extent.
- It will use some of the functionalities.
- Better than no coupling (fetch the data)
- It is suitable for small data sets.

(3) Semitight Coupling: 70%
- linked to the DB but not completely
- and some of the DM task primitives are also implemented in DB/DW

(4) Tight Coupling: 100%
- DM system is completely linked to DB/DW.
- It is most efficient among all schemas.
- The DB/DW system is fully integrated in such a way that it becomes part of the DM System.

- It is efficient and optimised implementation of DM.

Major Issues in Data Mining :- (or) DMs

1) Mining different kinds of knowledge in database.
   - In DB or Data Mining System it has many users and each user has their own interest or need. So based on the interest, it is required to mine different kinds of knowledge. Hence it is necessary for DM system to cover broad range of knowledge.

2) Interactive Mining of knowledge at multiple levels of abstraction
   - The data mining process needs to be interactive because it allows users to focus the search for patterns, providing & refining data mining requests based on returned results.

   Eg:- fetch rollnos → CSE and S and female

3) Incorporation of background knowledge
   - To guide discovery process and to express the discovery pattern, the background knowledge of the domain can be used.
   - The background knowledge may be used to

Express the discovered pattern not only in Concise terms but at multiple level of abstraction

4) Presentation and Visualization of data Mining results:

- Once the patterns are discovered it needs to be expressed in visual representation.
- This representation should be easily understandable by the users.

5) Handling noisy or incomplete data

- The data cleaning methods are required to handle noise, ~~irrel~~ incomplete data.
- If there are no data cleaning methods then the accuracy of the discovered patterns will be poor.

6) Efficiency and Scalability of data mining algorithm

- In order to extract effectively the information from huge amount of data in DB The DM algorithm must be efficient & scalable

Data Preprocessing :-

The process of transforming raw data into an understandable format.

Eg:- Names of students + Mark of students + Visual (raw data) (understandable for)

4-major tasks in data Preprocessing:

(1) data cleaning (3) Data Reduction
(2) data Integration (4) Data Transformation

(1) Data Cleaning: The process of removal of incorrect, incomplete, in accurate data, missing data is a data cleaning. also replace

There are two things in Data cleaning:

(i) Handling missing values/data.
(ii) Handling noisy data/values

Handling Missing Data: In place of missing values, replace with 'NA' (Not applicable), with mean value, with median value

in case of Normal distribution replace with mean

in case of Non-Normal dis replace with median

- Sometimes replace with most probable value (frequent value)
- Missing value can be filled in 2-ways
  (1) Manual — Small data sets
  (2) Automatic — large data sets; more efficient

Handling Noisy data : Noisy data is nothing but inconsistent / error data.

Methods to handle noisy data — There are 3-methods

(1) Binning (2) ~~Classification~~ Clustering (3) Regression

Binning :- Binning method smooth a sorted data value by consulting its neighbourhood value. Then the sorted data is stored in buckets or bins

- Once the data get stored into the bins perform Smoothing process which can be done in 3-ways removing error values (or) replacing

3-methods to handle data in bins :

(1) Smoothing by bin mean —

The values which are present in bin are replaced by bin value mean value

Eg :- 2, 3, 4, 5

Avg of bin = $\frac{2+3+4+5}{4} = 3.5$ } replace

∴ 3.5 will get replaced in bin by all values.

(2) Smoothing by bin median —

The values which are present in bin are replaced by median value.

Eg:- 1, 2, 3, 4, 5,
median.
∴ 3 will get replaced in bin by all values.

(3) Smoothing by bin boundary — min & max values
The values which are present in bin are replaced by min & max values.

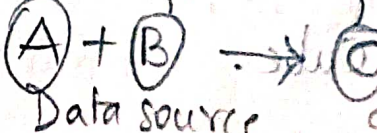
Regression:- Data Smoothing can also be done by Regression. This technique is used for the Numeric prediction of data.

Clustering:- This is also a method for data smoothing. In this technique, the similar data items are grouped at one place and dissimilar items fall outside the cluster. (outlier's)

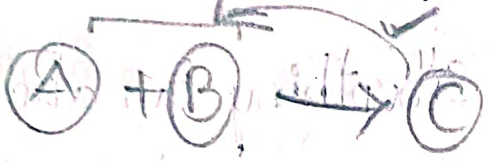
2) Data Integration: The process of combining multiple hetereogeneous sources of data into single dataset.

2-types of data Integration:-

(1) Tight Coupling — In this data is combined together into a physical location. Once the data is integrated it cannot be access separately.



(2) Loose Coupling – In this only an interface is created and data is combined through that interface and also accessed through interface.
- Data remains in actual database only.
- In this context, we can have access to individual datab& combine data base



(3) Data Reduction: Technique can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintain the integrity of the original data.
(Volume of data is reduced to make analysis easier)

Volume ↑ – Performance ↓,
Volume ↓ – Performance ↑

Eg: Size Comprehension – online converter

Methods for Data Reduction:
i) Dimensionality Reduction – In this, it reduces no. of input variables in the data set because large input variables has poor performance.
ii) Data Cube Aggregation – Data is combined to construct a data cube. Raw data

- In this, Redundant, noisy data is removed and a unique data cube is generated.

iii) Attribute Subset Selection (column) — In this, highly relevant attributes should be used and others should be discarded (removed). So in this case the data get reduced.

iv) Numerosity Reduction — In this, we store only a model/sample of data instead of entire data.

4) Data Transformation: It is a process where the data get transformed into appropriate form suitable for mining process

Methods for Data Transformation

i) Normalization — Normalization is done in order to scale the data values in specified range (-1.0 to 0.1 or 0 to 1)

- but in case of Name/Section it is not possible for scaling.

ii) Attribute Selection — In the Attribute Selection, new attributes are created using older ones (older attributes)

iii) Discretization — In this, raw values are replaced by interval levels.

Eg① age → raw value (10 yrs, 11 yrs, 21 yrs)
interval levels — 0-10, 11-20, 21-30,

iv) Concept hierarchy Generation — In this, attributes are converted from low level to high level.

Eg:- city → country
(low level) (high level)
attribute attribute

(3) Data Reduction:

Methods for data Reduction —

v) Wavelet Transforms — The discrete wavelet transform (DWT) is a linear signal processing technique, that applied to a data vector X, which transforms it into a numerical different vector X' of wavelet coefficient. X → X

* The two vectors are of same length (X, X')

* When applying this technique to data Reduction, we consider n-dimensional data tuple, i.e.

$X = (x_1, x_2, \ldots x_n)$

where 'n' is no. of attributes present in Data set

Data Reduction :-

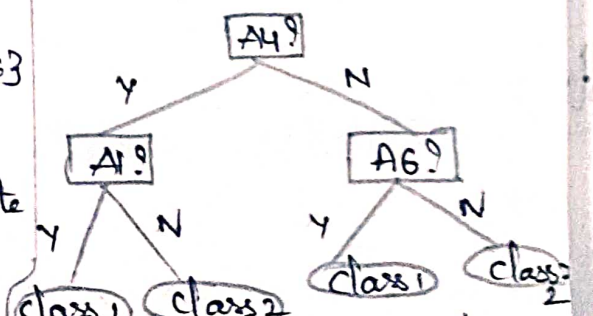
iii) Attribute Subset Selection — Reduces the data set size by removing irrelevant attributes.

The following are the techniques for attribute subset selection.

1) Stepwise forward selection — The procedure starts with an empty set of attributes as the reduced set. The best of the original attributes is determined & added to the reduced set. At each subsequent iteration or step are performed.

2) Stepwise backward Elimination — The procedure starts with the full set of attributes. At each step, it removes the ~~worst~~ worst attribute remaining in the set.

3) Combination of forward selection & backward elimination — These two methods can be combined so that, at each step, the procedure selects the best attribute & removes the worst from among the remaining attributes.

Forward Selection	Backward Elimination	Decision tree Induction
Initial Attribute set: {A1, A2, A3, A4, A5, A6}	Initial attribute set: {A1, A2, A3, A4, A5, A6}	Initial attribute set: {A1, A2, A3, A4, A5, A6}
Initial reduced set: { }	⇒ {A1, A3, A4, A5, A6}	
⇒ {A1}	⇒ {A1, A4, A5, A6}	
⇒ {A1, A4}	⇒ reduced attribute set:	
⇒ Reduced attribute set:	{A1, A4, A6}	
{A1, A4, A6}		

⇒ reduced attribute set: {A1, A4, A6}.

* The Wavelet transforms [illegible] data can be truncated & this is useful in data Reduction..
* This technique can also be used to remove the noise in the data..
* Wavelet Transforms also make the data cleaning very effective..

vi) Principal Component Analysis — PCA is a data Reduction technique that transforms/convert a large no. of correlated variable into a Smaller set of correlated variables called Principal components
* It is mainly used for the dimensionality reduction technique in AI, Computer Vision & Image Compression.